The authors demonstrate how item response theory can be applied for cost effective measurement of consumer attitudes with multi-item scales. The measurement technique they discuss and illustrate is tailored to each respondent so that each is asked only the scale items most informative of his or her attitude level. This approach yields attitude estimates from only a fraction of the total number of items in the attitude scale, with a measurable and controllable increase in the standard error of measurement. Potential cost-saving implications are discussed.

Measuring Consumer Attitudes Toward the Marketplace With Tailored Interviews

A common dilemma facing survey researchers is the tradeoff between increasing the quantity of information gathered in an interview and the higher interviewing costs incurred as a result. These costs include not only the time costs directly stemming from administering interviews of increased length, but also the potential measurement errors due to respondent fatigue, decreased involvement, and other factors.

We advocate a “tailored” interview process, a measurement approach that decreases interviewing costs (by using only a fraction of the number of items in a multi-item measurement scale) while controlling measurement accuracy. We illustrate this approach empirically in the context of measuring consumer attitudes toward the marketplace, using two scales available in the marketing literature.

In this tailored interview approach to attitude measurement, the key to decreasing interviewing costs is obtaining an a priori attitude estimate of the respondent based on a few scale items and selecting additional scale items that would provide the best update of this prior attitude measure. Note that any attitude scale consists of several items, each of which is most informative at a unique point on the attitude continuum. Therefore, if a person has a highly negative attitude, measurement can be made more efficient by asking only those scale items that are most informative at the high negative end of the attitude continuum. Consider, for example, the following subset of four items drawn from a hypothetical scale to measure consumer attitudes toward unethical business practices, with dichotomous (agree/disagree) response options.

1. A business that resorts to any unethical practice must be severely punished.
2. Successful business practices and high ethical standards can go together.
3. Unethical business practices of a minor nature should be condoned.
4. Each manager in each business must be held personally responsible for any unethical practices within his or her department.

If a respondent agrees with items 1 and 2 (demonstrating a high negative attitude), a response to item 3 (a moderately positive attitude) will yield little additional information because one can predict that the respondent would disagree with it. In other words, item 3 will provide only redundant information (that the respondent does not carry a moderately positive attitude) and can be excluded from the interview session. In contrast, it seems worthwhile to add item 4 to the interview because it might indicate how highly negative the respondent’s attitude is.

In sum, the process involves “tailoring” the measurement session to the individual in such a way that the sequence of items asked in the interview depends on the responses to items asked before. It is useful to delineate several differences between this approach and the traditional method of scale measurement. First, the theoretical foundations of these approaches are markedly dif-
different: the former is based on item response theory (or IRT; also known as latent trait theory) whereas the latter is derived from classical measurement theory (or CMT; also known as the ‘traditional’ measurement approach).

Second, the two approaches differ markedly in their orientation toward assessing measurement error. The traditional approach relies on interitem and item-total correlations for assessing measurement error. These correlations (and other reliability indices such as Cronbach’s alpha) are computed over the entire attitude range; further, the measurement error is assumed to be constant over all attitude levels. In contrast, IRT models assess measurement error at specific attitude levels. Moreover, the addition of items to the measurement scale that are similar (i.e., tapping the same attitude range) will improve reliability as defined by the traditional methods, even though it may not contribute to better measurement at other levels in the attitude continuum. In other words, reliability coefficients do not distinguish between redundant items tapping the same attitude levels and distinct items tapping different levels of the same attitude continuum. In contrast, in the IRT approach, if such additional items are not informative at the attitude level of a given respondent, they will carry little relevancy for determining his or her attitude level.

Finally, to allow comparisons among respondents on the attitude being measured, the CMT approach requires all respondents to answer all items in the attitude scale. However, because IRT-based attitude measurement is independent of the particular items used in the interview (provided some items overlap across respondents), attitude measures are comparable even if they are based on different items. This IRT property facilitates the idea of tailoring the items in an interview to the respondent’s attitude structure as the interview proceeds. Note that such tailoring cannot be accomplished under CMT because that approach uses summed scores requiring responses to all scale items.

In the following three sections we present a brief overview of item response theory and the tailored interview procedure, report an empirical application of this measurement technique in the context of assessing consumer attitudes toward the marketplace, and derive conclusions from the study.

**ITEM RESPONSE THEORY AND THE TAILORED INTERVIEW PROCEDURE**

This review provides a brief background on IRT and the tailored interview procedure. The interested reader is referred to several landmark contributions to IRT for more details (Birnbaum 1968; Bock 1972; Hambleton et al. 1978; Hambleton and Swaminathan 1985; Hulin, Drasgow, and Parsons 1983; Lazarsfeld 1950, 1959; Lord 1980; Rasch 1980; Samejima 1969; Wright and Mead 1977; Wright and Panchapakesan 1969). In particular, the educational measurement literature has devoted the most attention to IRT-based models that tailor items in an ability test to the particular individual being measured (e.g., Green 1983; Ury 1977; Weiss 1982). IRT-related applications in the marketing literature include those of Kamakura and Srinivasan (1983) and Bechtle (1985).

It is worthwhile to begin the description of IRT models with a brief discussion of their relationship to the Guttman scale. Both Guttman and IRT models assume that items and respondents can be ordered along the same unidimensional trait continuum. The basic difference is that the former specifies a deterministic relationship between item responses and the individual’s trait whereas the latter assumes that the probability of responding to an item positively (a positive answer is defined as an item response that suggests the individual has the trait being measured) is a function of the individual’s trait level and the item’s characteristics; this probability is often postulated as a logistic function of item parameters and the individual’s trait.

**Birnbaum’s Two-Parameter Logistic Model**

In particular, we direct our attention to Birnbaum’s (1968) two-parameter logistic model given by:

\[
P_i(\theta_j) = \frac{1}{1 + \exp\{-Da_i(\theta_j - b_i)\}}
\]

where:

- \(\theta_j\) = trait/attitude level \(j\),
- \(P_i(\theta_j)\) = probability of responding positively to item \(i\) at \(\theta_j\),
- \(a_i\) = discrimination parameter for item \(i\),
- \(b_i\) = threshold or “difficulty” parameter for item \(i\), and
- \(D\) = scaling constant (equal to 1.7).

This function, also known as the item characteristic curve (ICC), is depicted in Figure 1A. The hypothetical items 1, 2, and 3 shown in the figure have discrimination parameters \(a_1 = 1.0\), \(a_2 = 1.5\), and \(a_3 = .8\), respectively, and threshold parameters \(b_1 = 1.0\), \(b_2 = .5\), and \(b_3 = -.5\), respectively. The discrimination parameter \(a\) is a function of the slope of the ICC at its inflection point (where \(P_i(\theta_j) = .50\)) and the threshold parameter \(b\) corresponds to the trait value \(\theta\) at the inflection point of the ICC (Hulin, Drasgow, and Parsons 1983). Clearly, the ICC for the high discrimination item 2 is the steepest, indicating that the probability of positive response to this item is highly sensitive to the variation of traits in the neighborhood of the inflection point. Accordingly, item 2 is highly capable of discriminating between traits.

---

1 The term “trait” is perhaps the most frequently used descriptor in the measurement literature of the construct being measured. Though we use this term largely for the reason above, our description of IRT/CMT applies to the measurement of traits as well as to the measurement of attitudes.

2 Birnbaum (1968) has suggested the inclusion of a third parameter \(c\), to account for the probability that low-trait subjects would answer an item correctly by pure guessing.
just above or below the trait level \( b = .5 \); a positive response to this item would indicate a strong likelihood that the respondent’s trait level is above \( b = .5 \) whereas a negative response would indicate the opposite. Therefore, if we had an \textit{a priori} expectation that an individual’s trait is in the vicinity of \( b = .5 \), item 2 would be the most useful item in reducing the uncertainty about his or her trait level.

The amount of information provided by the item at a particular trait level can be computed by

\[
I_i(\theta) = \frac{(P_i'(\theta))^2}{P_i(\theta)(1 - P_i(\theta))}
\]

where \( P_i'(\theta) \) is the first derivative of equation 1 with respect to \( \theta \).

The information function computed from equation 2 is related inversely to the asymptotic variance of the trait estimate obtained with the item at the trait level \( \theta_i \) (Hambleton and Swaminathan 1985). Therefore, this function serves as an important criterion for the selection of items given a prior estimate of the subject’s trait level.

The information functions corresponding to each of the three items discussed before are plotted in Figure 1B, from which two points are observed. First, if we compare information levels for \textit{any given item}, clearly each item is most informative at its threshold level \( b_i \). Second, if we compare information levels \textit{across items}, item 2 provides the highest level of information. This observation is in line with what one would expect, because item 2 has the highest discrimination parameter.

Though we employ the two-parameter logistic model here, the literature contains more complex extensions of IRT models to accommodate multichotomous items (Bock 1972), items scored in graded responses (Anderson 1977; Andrich 1978; Samejima 1969), continuous-response items (Samejima 1973), and multidimensional traits (Mulai 1972; Reckase and McKelvie 1982, 1985; Samejima 1974). Discussion of these models is beyond the scope of our article.

\section*{Item-Free Estimation: An IRT Characteristic}

The most important IRT characteristic in the context of tailored interviewing is \textit{item-free estimation} of traits. That is, the IRT estimate of a respondent’s trait is independent of the particular set of items (from the scale measuring the trait) that he or she answered. When an IRT model fits data on the measurement scale of interest, the trait estimate obtained for each respondent is an asymptotically unbiased measure of the true trait level, irrespective of the number of items administered (Hambleton and Swaminathan 1985). A direct implication of this IRT property is that the trait estimates of two or more respondents will be comparable even if those respondents answered different sets of items measuring the same unidimensional trait. Obviously, this view sharply contrasts with the CMT approach, which holds that the trait measure is a linear combination of item responses and that the measures of two individuals on a given trait are comparable only if they were based on the same or a parallel set of items.

\section*{Tailored Interviewing Process}

The first requirement for conducting tailored interviews is an item bank—a set of measurement scale items with estimated threshold (\( b_i \)) and discrimination (\( a_i \)) parameters. These parameter values for each item are estimated by using response data from a “calibration sample” of respondents drawn from the same population as that addressed in subsequent tailored interviews.\(^3\) Detailed discussions of the procedures for estimating item parameters are available in several sources (Bock 1972; Lord 1974; McKinley and Reckase 1980; Wright and Douglas 1977) and the researcher can choose one of many mainframe computer programs for the estimation task (LOGOG, Kolakowski and Bock 1973; ANCILLIES, Ury 1978; LOGISTICS, Wingersky, Barton, and Lord 1982; BILOG, Mislevy and Bock 1982).

The basic issues in developing a tailored interviewing procedure are:

- which item (or items) to ask first,
- given the answer(s) to the first item(s), which items to ask next,
- how to compute the trait estimates based on the observed responses, and
- when to stop the interview

It is appropriate at the outset of the tailored interview to ask a few items with widely different threshold values (i.e., items representing different levels of the trait continuum). This strategy is useful because it provides a rough indication of where the respondent belongs on the trait continuum. Next, as discussed before, the information function is a useful criterion in determining the sequence of items to be asked in the interview. To address the third issue, the adaptive testing literature in the field of educational measurement suggests two broad approaches: maximum likelihood estimation (MLE) of logistic models (Birnbaum 1968) and a Bayesian updating approach for normal ogive models (Jensen 1974; Owen 1969, 1975). On balance, the MLE approach appears more advantageous because it is known to provide asymptotically unbiased and efficient results whereas the Bayesian approach may lead to biased estimates under some conditions (Weiss and Kingsbury 1984). More-

\(^3\)In practical adaptive (or “tailored”) testing situations, this is usually done by administering all of the scale items to a respondent sample that is suitably large and comparable (Green et al. 1984) to the sample that will be interviewed later. It is useful to note that the calibrated item parameter values (for any IRT model) are independent of the particular calibration sample used as long as this sample is drawn from the population of measurement interest (see Hambleton and Swaminathan 1985, p. 11–13). The one-parameter logistic model (also called the Rasch model) is a special case, when using this model, one can calibrate items with any sample respondents drawn from any population (see Masters 1982).
over, procedures are readily available for the MLE of
dichotomous response (Birnbaum 1968; Wood, Winger-
sky, and Lord 1976) and multichotomous response models
(Bock 1972; Samejima 1969). We therefore use the MLE
approach. Details on the MLE of traits based on the
two-parameter logistic model (equation 1) are presented
in the Appendix.

The MLE procedure assumes that an individual's responses are
independent over items. Further, it yields asymptotically unbiased trait
estimates, unless all items are answered positively (or all negatively)
in every step of the process. (In such cases, the MLE procedure will
be unable to provide an estimate of the individual’s trait level; this
drawback also applies, however, to any trait measurement procedure
unless Bayesian methods of estimation are used.)

The final issue is when to stop the interview. Jensen (1974) suggests two stopping rules for the tailored pro-
cedure: stopping when (1) a prespecified minimum stan-
dard error of measurement is reached for each respon-
dent or (2) when a maximum number of items are
answered by each respondent. In the empirical applica-
tion described in the following section, we employ
both rules.

MEASURING CONSUMER ATTITUDES TOWARD
THE MARKETPLACE

Much has been written about consumerism and about
how businesses have responded to it (Aaker and Day 1972;
Day and Aaker 1970; Dembinski 1980; Greyser and Dia-
mond 1974; Kotler 1972; Peterson 1974). This focus on

To illustrate empirically the tailored interview procedure, we have chosen to focus on the measurement of consumer alienation and consumer discontent for several reasons. First, these constructs represent consumer attitude domains that are broad and general; they include all facets of business and the marketplace that are meaningful to the consumer. In contrast, consumer satisfaction/dissatisfaction and consumer complaint behavior represent narrower domains in that the focus is on attitudes toward specific products and/or businesses.

Second, well-established measurement scales for the alienation (Allison 1978) and discontent (Lundstrom and Lamont 1976) constructs are available in the literature. Third, on theoretical grounds (Allison 1978; Lambert 1980) and from past empirical evidence (Lundstrom, Kerin, and Sciglimpaglia 1979; Singh 1985), we can expect these two constructs to be substantially correlated. This strong interrelationship is important for our analyses because it provides an opportunity for the cross-validation (or external validation) of the measures of these two constructs. Finally, the study of consumer aliena-
tion/discontent is important because it can provide crucial inputs for modifying marketing strategies, programs, and practices in order to protect business interests (Lambert 1980). Lambert and Kniffen (1975) cite evidence that could sensitize business to the dangers of ignoring increased alienation among consumers; further, they suggest that business sensitivity to alienation in consumers can yield better strategies to mitigate consumer discontent.

A brief synopsis of each of these two constructs and corresponding measurement scales follows.

—Consumer alienation reflects the "feelings of separation from the norms and values of the marketplace: lack of acceptance of or identification with market institutions, practices, and outputs as well as feelings of the self when one is involved in the consumption role" (Allison 1978, p. 570). This scale contains 35 items. Reliability and validity estimates of the scale are given by Allison (1978).

—Consumer discontent captures the attitude toward "the product strategies of business, business communications and information, the impersonal nature of business and retail institutions, and broader socio-economic forces which are linked with the economic system" (Lundstrom and Lamont 1976, p. 374). This scale consists of 82 items designed to measure a consumer's attitude toward marketing practices rather than toward a specific product or firm. Though originally 173 statements were created to include various market-related factors, reliability and validity estimates of the refined scale showed strong internal consistency, homogeneity, and external validity (Lundstrom and Lamont 1976).

An important constraint to the widespread and routine application of these two scales is the large number of items required by both. Paper-and-pencil-based administration of these two scales (totaling 35 + 82 = 117 items) generally requires more than 30 minutes for the average respondent. However, our empirical tests show that considerable time savings can be achieved via tailored interviews because substantially fewer items are administered in comparison with the two full scales. More importantly, the time savings do not come at a heavy price; we show that tailored interviews lead to a small and controllable loss in measurement accuracy.

Scale Calibration

The 35 items of the alienation scale and 82 items of the discontent scale were administered to a convenience sample of 541 adults in a midwestern city via a paper-and-pencil task. We used the data from this calibration sample to estimate the item parameters for all items belonging to each of these two scales, thereby creating two (scale-specific) item banks to be used in the tailored interviews. For both scales, a chi-square test could not reject the hypothesis that the estimated model fit the observed frequency distribution of the answers (at α = .21 and .29 for the alienation and discontent scales, respectively).

The item threshold (b) and discrimination (a) parameters for all items in the alienation and discontent scales are plotted in Figures 2 and 3, respectively (each item is identified by its sequence number). Given that the threshold parameter b, for any item is equal to the trait estimate θ at the inflection point on the ICC; both scales clearly cover a reasonable range of the attitude continuum (the θ measures are standardized, thus concentrating in the −2/+2 range), though the discontent scale covers a broader range.

Figures 2 and 3 also show that the highly discriminating items in both scales are characterized by moderate threshold levels, that is, they are concentrated in attitude levels in the middle of the attitude continuum. From equations 1 and 2, it follows that the highest amount of information provided by an item with discrimination a is I(θmax) = D′a2/4 (see Lord 1980, p. 151).

Therefore, the maximum information provided by each item is related directly to the parameter a. Figures 2 and 3 thus show that the full-scale (i.e., all scale items administered) measures are more informative at the middle attitude levels, which are dominated by high discrimination items. As a result, application of the full scale to measure consumer attitudes would lead to more accurate measurements at midrange than at the two extremes. Reliability measures, in contrast, would indicate only the overall average accuracy of the scale, ignoring this variation across attitude levels.

Implementation of the Tailored Interview

The a, and b, values for the various items of the alienation and discontent scales are represented by their respective locations in the parameter space of Figures 2 and 3, respectively. For each scale, the estimated parameter values for all scale items constituted an item bank from which items were drawn during a tailored interview pertaining to that scale. In other words, the data in Figures 2 and 3 were the item banks for alienation and discontent scales, respectively. Given the item bank for the scale of interest, the tailored interviewing algorithm proceeded along the following steps.

A BASIC program (PCTAILOR) was developed to implement the tailored process on a personal computer. PCTAILOR first asked three items covering a broad range of threshold levels: one item each from the low end, middle, and high end of the attitude continuum. A pre-

5The adaptive testing literature suggests several criteria for assessing the suitability of an item bank for tailored testing. For instance, Ury (1977) suggests that (1) item threshold (b) values must be in the −2 to +2 range and (2) item discrimination (a) values must be 8 or higher. The item banks for the alienation and discontent scales met or exceeded the requirements of criterion 1. Further, most of the items in both these scales also satisfied criterion 2, to preserve the integrity of the measurement scales as originally developed, we did not drop items that failed this requirement. Moreover, Green et al. (1984, p. 350) recommend against rigid enforcement of criterion 2.

6When PCTAILOR presented a given scale item, the respondent answered the item on the basis of a symmetrical 6-point Likert scale.
Figure 2
ITEM PARAMETERS, ALIENATION SCALE

![Graph](image)

Preliminary maximum likelihood (ML) estimate of the respondent's attitude then was obtained from the answers to these items by means of equation A2 in the Appendix. In other words, the program used the responses to these initial items and the fixed \( a \) and \( b \) values from the item bank to compute the preliminary ML estimate (if all three items were answered in the same direction, an ML estimate was not possible and an additional item was asked). Given this preliminary attitude measure (\( \hat{\theta} \)), the item bank parameters, and equation 2, PCTAILOR computed the amount of information expected of each item (from the pool of items not used so far), selected the most informative item at this preliminary attitude measure level, and presented it to the respondent. The program then revised the attitude measure (obtained before) on the basis of the answer to the new item by computing a new ML estimate of the respondent's attitude; this process of updating the previously obtained attitude estimates was repeated until a prespecified stopping rule was satisfied. At each step, PCTAILOR stored the attitude measure obtained, the standard error of measurement, and the time elapsed since the beginning of the interview.

In the interviews described next, no stopping rule was applied and hence all items were asked in the order defined by the tailored interview. This procedure made
possible tests of the internal consistency and external validity of the tailored-interview measurements at various stages in the interview. Nevertheless, each interview was conducted according to the tailored procedure; attitude estimates at any given point were tailored estimates, whereas the estimates after the last item were IRT-based full-scale measurements.

Tailored interview sessions involving both the alienation and discontent scales were administered via PCTAILOR to a sample of 137 college students (business, engineering, computer science, and psychology majors) in a midwestern university. All of the following analyses relate to the data gathered from this interview sample.

Sample Tailored Interview Session

Figure 4 depicts the results from a sample tailored interview session (covering the discontent scale). The data are representative of those obtained from other respondents. The chart displays the attitude estimate after each item, along with a 95% confidence interval.

After the three initial items, this person's attitude measure was $\theta = 1.05$ (a somewhat discontented subject, because high $\theta$ levels imply discontentment). The answers to the next few items (chosen on the basis of how informative they were at each successive attitude level estimate) led to a decrease and then an increase in the attitude estimate. Nevertheless, the attitude measure for this person was fairly constant after 20 items, at $\theta = 1.0$. Furthermore, the confidence interval (or the standard error of measurement) remained virtually unchanged after the first 20 items.

These results graphically exemplify the marginal improvement in accuracy achieved by asking several of the remaining (less informative) items in the later part of the tailored interview. They suggest that potential time and
cost savings could have been realized by terminating the tailored interview at some point before all 82 items in the discontent scale were asked. More convincing results, obtained across the entire sample of 137 respondents, are presented next.

**Aggregate Empirical Results**

Though the tailored approach was used to measure attitudes in the interview sample, each respondent was required to answer all items in each scale. This research design feature enabled us to compare the results at different stages of the tailored interview with the results from administration of the full scales and also provided some insights about the accuracy and validity of the measurements obtained.

Figure 5 summarizes the relationship between the standard error of the attitude measurement and the proportion of the total interview time (within each scale) needed to achieve it. These curves represent bilog regressions estimated across all respondents and over all interviewing stages, with $R^2$ values of .78 and .91 for the alienation and discontent scales, respectively. They show the most substantial improvement in accuracy at the earlier stages of the tailored interview (when the most informative or most suitable items are asked for each respondent) and the marginal improvement observed at the later stages (after 30% of the interview time has already been spent).

According to these curves, the standard error of mea-
measurement would not be substantially affected if the interviews were interrupted after about 30% of the total interview was conducted (for each scale). In other words, by using the tailored interview process, we could have decreased interviewing time by 70% with no major impact on the measurement accuracy.

Figure 5 is a plot of the IRT-based attitude estimates (thetas) and the standard error of measurement for each of the 137 respondents in the interview sample, which were obtained with the full administration of the two scales. This figure confirms our expectations based on the item parameters in Figures 2 and 3; both scales lead to more accurate measurements at the middle of the attitude range. Further, this result contradicts the usual assumption of constant measurement error (homoscedasticity) implied by reliability coefficients. Figure 6 also suggests that a smaller number of items might be needed to measure attitudes at the midrange with measurement error comparable to the error attained at the more extreme attitude levels where more items would be administered.

Figure 7 shows the results obtained by stopping the interview (i.e., using only the response data up to a given point) when the standard error of measurement reached 2 or less. This stopping rule would lead to a comparable measurement error for all respondents over the whole attitude continuum. The figure displays the proportion of total interview time (for each scale) required to achieve the minimum acceptable measurement error (2) at each attitude level.

The results for the discontent scale show that, with the exception of a few cases of extremely low discontent (θ, \( < -2 \)), the specified measurement error was attained after using only between 40 and 55% of the total interview
time, representing a substantial reduction of interviewing time, cost, and burden on the respondent. The results for the alienation scale (which had only 35 items), in contrast, show that the full scale was used for attitude levels lower than −1.0 (though not necessarily reaching the specified accuracy of minimum .2 standard error) and 70% of the total interview time was required at the middle range to achieve the desired accuracy level.

The general conclusion from these results is that tailoring the measurement instrument to the attitude level of each respondent may lead to substantial reductions in interviewing time and costs, depending on the characteristics of the item bank. For example, if the item banks had contained more items that were highly informative at the extreme attitude levels, the standard errors would have been lower for respondents with extreme attitudes.

The results presented so far indicate the cost-accuracy tradeoffs in terms of the expected measurement error (standard error of the attitude estimate) rather than direct empirical observation of measurement error. Obviously, the latter is not possible because “true” attitude scores are not observable. Therefore, we attempted to evaluate the tailored process by computing correlations between measurements obtained from subsets of items (from the tailored interview) and the IRT measurements obtained from the two full scales, which were taken as (imperfect) proxies for the “true” attitudes.

In making this assessment, we are assuming that the IRT-based full-scale measure is the closest to the “true” attitude though in effect this full-scale measure is also affected by measurement error. Because both measures of interest (the full-scale and tailored-interview results) are affected by measurement error, one should expect some attenuation in the correlations between measurement...
ments. This attenuation will be stronger the higher the measurement error Therefore, the correlations between measurements reported hereafter are underestimates of the strength of the actual relationships and should be viewed as lower bounds for the "true" correlations.

The first two columns of Table 1 contain the correlations between estimates derived from IRT. The correlations between the measurements based on tailored subsets and the IRT-based full scale are fairly high (.93) after 20 and 30 items for the alienation and discontent scales, respectively. These strong correlations demonstrate that little information is lost by administering only a limited subset (of the full scale) containing the "best" items. Because the interview was conducted so that the most suitable item (in terms of the amount of information about the respondent's attitude) was asked at each stage, fairly consistent attitude measurements are obtained with only a fraction of the item bank. Though these results demonstrate strong consistency with the IRT-based full-scale measurements, they give no information about the validity of the tailored-interview measurements.

Fortunately, there are theoretical grounds as well as prior empirical evidence for expecting strong correlations between the alienation and discontent constructs. Therefore, the correlation of the measurements of the two constructs provides useful insights to the (external) validity of both measurements.

The cross-correlations between tailored estimates of alienation and the full-scale IRT estimate of discontent (Table 1) are fairly stable after 20 items. The external validity of the tailored measurement of alienation (with the full-scale IRT measurement of discontent as the external criterion) remains virtually the same whether 20
Table 1
CORRELATIONS OF ATTITUDE ESTIMATES ACROSS 137 RESPONDENTS*

<table>
<thead>
<tr>
<th>Attitude</th>
<th>IRT estimates</th>
<th></th>
<th>CMT estimates</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alienation</td>
<td>Discontent</td>
<td>Alienation</td>
<td>Discontent</td>
</tr>
<tr>
<td></td>
<td>All 35 items</td>
<td>All 82 items</td>
<td>Sum of 35 items</td>
<td>Sum of 82 items</td>
</tr>
<tr>
<td>Alienation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRT/tailored 10 items</td>
<td>72</td>
<td>60</td>
<td>54</td>
<td>45</td>
</tr>
<tr>
<td>IRT/tailored 15 items</td>
<td>83</td>
<td>68</td>
<td>67</td>
<td>53</td>
</tr>
<tr>
<td>IRT/tailored 20 items</td>
<td>93</td>
<td>74</td>
<td>76</td>
<td>59</td>
</tr>
<tr>
<td>IRT/tailored 25 items</td>
<td>97</td>
<td>77</td>
<td>82</td>
<td>65</td>
</tr>
<tr>
<td>IRT/tailored 30 items</td>
<td>99</td>
<td>76</td>
<td>85</td>
<td>65</td>
</tr>
<tr>
<td>IRT/all 35 items</td>
<td>100</td>
<td>77</td>
<td>86</td>
<td>66</td>
</tr>
<tr>
<td>CMT/sum of 35 items</td>
<td>86</td>
<td>67</td>
<td>1.00</td>
<td>69</td>
</tr>
<tr>
<td>Discontent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRT/tailored 20 items</td>
<td>71</td>
<td>.83</td>
<td>.65</td>
<td>71</td>
</tr>
<tr>
<td>IRT/tailored 25 items</td>
<td>.76</td>
<td>90</td>
<td>.68</td>
<td>76</td>
</tr>
<tr>
<td>IRT/tailored 30 items</td>
<td>.77</td>
<td>93</td>
<td>.70</td>
<td>81</td>
</tr>
<tr>
<td>IRT/tailored 35 items</td>
<td>77</td>
<td>95</td>
<td>69</td>
<td>82</td>
</tr>
<tr>
<td>IRT/tailored 40 items</td>
<td>77</td>
<td>97</td>
<td>68</td>
<td>81</td>
</tr>
<tr>
<td>IRT/tailored 45 items</td>
<td>77</td>
<td>98</td>
<td>68</td>
<td>80</td>
</tr>
<tr>
<td>IRT/tailored 50 items</td>
<td>77</td>
<td>98</td>
<td>68</td>
<td>.78</td>
</tr>
<tr>
<td>IRT/tailored 60 items</td>
<td>77</td>
<td>99</td>
<td>67</td>
<td>80</td>
</tr>
<tr>
<td>IRT/tailored 70 items</td>
<td>77</td>
<td>1.00</td>
<td>67</td>
<td>80</td>
</tr>
<tr>
<td>IRT/all 82 items</td>
<td>77</td>
<td>1.00</td>
<td>67</td>
<td>80</td>
</tr>
<tr>
<td>CMT/sum of 82 items</td>
<td>66</td>
<td>.80</td>
<td>69</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*IRT = item response theory, CMT = classical measurement theory
All correlations significant at .001 level Figures in italics represent cross-correlations (for external validation)

items ($r_{AD} = .74$) or all 35 items ($r_{AD} = .77$) are used. Similar conclusions can be drawn for the tailored measurements of discontent with the full-scale IRT measurement of alienation as the external criterion.

Because the tailored interviewing measurements are based on a subset of the full scales, one might expect losses in validity. However, our external validity tests using the discontent and alienation constructs for cross-validation show no substantial loss.

Note that the preceding discussion is limited to correlations between tailored estimates and IRT-based full-scale estimates. Because both the alienation and discontent scales were developed on the basis of classical item analysis (using a Likert format) and because these constructs often have been measured as simple summed scores (Allison 1978; Lundstrom, Kerin, and Sciglioppa 1979; Lundstrom and Lamont 1976), it would be informative to investigate the correlations of the tailored estimates with CMT-based summed scores as well. These correlations are listed in the third and fourth columns of Table 1. In general, the conclusions drawn before from the first and second columns also apply for the third and fourth columns. However, the correlations between tailored estimates and the CMT-based summed scores are lower than the corresponding correlations with IRT estimates, implying higher measurement error in either one of the measurement methods. In contrast, the cross-correlation (between the two constructs) obtained from the CMT method (.69) is clearly lower (or more attenuated) than that obtained with the IRT method (.77), indicating better measurement accuracy for the latter.

In summary, the analyses of correlations in Table 1 reinforce our conclusion that the tailored approach to attitude measurement considerably enhances measurement efficiency.

**DISCUSSION AND CONCLUSIONS**

The "traditional" (or CMT) approach to developing and using consumer attitude scales reflects the philosophy "one size fits all." Attitude scales are constructed, validated, and applied in such a way that every respondent is subjected to the same battery of questions, regardless of his or her attitude level.

The IRT-based approach applied and tested in our study tailors the scale to the measurement requirements at each attitude level so that each respondent is subjected only to the items that are most informative about his or her specific attitude level, leading to time and cost savings and less burden on the respondent. Furthermore, this approach provides a means to assess the measurement error at each attitude level and enables the researcher to prespecify the maximum acceptable measurement error.

The main focus of our empirical tests was the assessment of accuracy and validity of tailored interviewing applied to the measurement of consumer alienation and discontent. Applications of unifactor IRT models (such as the two-parameter logistic model used here) presuppose unidimensionality in the construct measured (Ham-
Evidence supporting the undimensionality assumption for the alienation scale stems from previous research (Allison 1978, p. 569-70). Further, a strong Cronbach's alpha (.868) provides evidence of internal consistency for this scale.

The discontent scale was mainly useful for evaluating the external validity of the alienation measurements. However, application of the tailored interviewing technique to measure discontent is worth discussing. Though this scale is known (from factor analyses) to span several dimensions (Lundstrom and Lamont 1975, 1976), the discontent measures obtained via tailored interviews still appear to be very encouraging in terms of accuracy, potential cost/time savings, and external validation with the alienation measures. Additionally, this scale has a high alpha coefficient (.93), indicating high internal consistency.

The preceding results are in line with Drasgow and Parsons' (1983) suggestion that applications of unifactor IRT models need not be restricted to very narrow (i.e., strictly unidimensional) constructs, but can include a reasonably broad construct as long as it is characterized by a moderately potent general trait. Furthermore, Reckase's (1979) results indicate that unifactor IRT models, when applied to multidimensional constructs with a dominant factor (such as in our case), will yield trait measurements that are highly correlated with that dominant factor.

Though the application presented here involves data with dichotomous response categories, several IRT models are available in the literature that can accommodate multichotomous responses (Anderson 1977; Andrich 1978; Bock 1972; Samejima 1969) in tailored interviewing contexts. Similarly, if multidimensional constructs are to be measured, several latent trait models are also available (Mulaik 1972; Reckase and McKinley 1982, 1985; Samejima 1974; Symposion 1978). Multidimensional tailored testing procedures already have been tested (Urry 1977) with encouraging results.

Though the primary attraction of the tailored interview approach is the promise of potential cost reductions, several secondary reasons justify the application of this technique. The administration of the tailored algorithm in telephone surveys presupposes the availability of an interactive computer environment. With the growing popularity of the CRT interviewing technology (Tyebjee 1979), the availability of infrastructure for administering tailored interviews does not appear to be a problem. Further, with the rapid diffusion of personal computers, a large number of potential respondents are likely to have the minimal computer familiarity/skills required for the application of the tailored technique via direct respondent interaction with computers.

In conclusion, we note that prior empirical research addressing the measurement of intellectual ability in the educational measurement field provides supportive evidence of several advantages for the "tailored testing" approach, such as measurement accuracy and cost/time savings. Though these benefits may be of enormous potential value in market research and survey work, use of the tailored measurement technique has not been explored in the marketing literature. Our empirical work extends the principles of tailored testing to measure attitudes of relevance to marketers and indicates that the tailored interview approach offers cost/time saving without significant sacrifices in measurement accuracy. Further, we demonstrate that the application of the technique is appropriate even when the construct measured is reasonably broad, as long as it taps a general prepotent dimension. More important, our tailored interview measures are robust not only to within-scale comparisons with IRT-based and CMT-based full-scale estimates, but also to the more stringent criterion of external validation. Given the potential cost savings and other benefits offered by this technique, marketing researchers should vigorously pursue its widespread application.

APPENDIX

MAXIMUM LIKELIHOOD ESTIMATION OF A TRAIT BY THE TWO-PARAMETER LOGISTIC MODEL

The first-order condition for the MLE of an individual's trait \( \theta \), based on responses to \( k \) scale items, by the two-parameter logistic model is

\[
\frac{\partial l}{\partial \theta} = D \sum_{i=1}^{k} a_i [V_i - P_i(\theta)] = 0
\]

where,

- \( l = \) log-likelihood function for the observed set of \( k \) responses,
- \( V_i = 1 \) if item \( i \) is answered positively and 0 otherwise, and
- \( P_i(\theta) = \) the probability of answering item \( i \) positively. The terms \( P_i(\theta) \), \( a_i \), and \( D \) are defined in equation 1.

The solution for equation A1 can be obtained through Newton's gradient search. Let

\[
\theta_q = \text{trait estimate at step } q \text{ of the search } (\theta = 0 \text{ can be used as an initial estimate, assuming a median ability}).
\]

Then the estimate can be revised iteratively with:

\[
\theta_{q+1} = \theta_q + \Delta \theta_q
\]

where

\[
\Delta \theta_q = \frac{\partial l/\partial \theta_q}{\partial^2 l/\partial \theta_q^2}
\]

\[
\partial l/\partial \theta_q = D \sum_{i=1}^{k} a_i [V_i - P_i(\theta_q)]
\]

\[
\partial^2 l/\partial \theta_q^2 = -D^2 \sum_{i=1}^{k} a_i^2 [1 - P_i(\theta_q)] P_i(\theta_q)
\]
Note that this is a simple unidimensional search that can be performed very quickly.

An asymptotic estimate of the standard error $S$ of the trait estimate can be obtained as $S = 1/\sqrt{-\partial^2 / \partial \theta^2}$ where $\theta$ is the MLE obtained at the termination of the gradient search.

REFERENCES


